AN ANALYSIS OF THE PERFORMANCE OF ABOVE AVERAGE READERS AND BELOW AVERAGE READERS ON A PROGRAM-DEPENDENT MASTERY TEST

By

DARLA V. McCREA

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN
PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

1983

ACKNOWLEDGEMENTS

The writing of a dissertation is usually perceived as a solitary venture. This perception is false. The efforts of many people have enabled me to finish this project.

I wish to thank all my committee members: Dr. William Hedges, Dr. Linda Crocker, Dr. Gordon Lawrence, Dr. Sue Kinzer, and Dr. Ruthellen Crews. Their advice and guidance have been crucial to my development as a professional as well as to the completion of this project. Special thanks go to Dr. Hedges for being an advisor and friend for so long, to Dr. Crocker for the abundance of support and help during the past year, and to Dr. Lawrence for inspiration.

I wish to thank the Curriculum Resource Teachers who gathered the data for this study, and Gayle from the Research Department who assisted in matching student numbers.

Thank you to very dear friends, Beth, Mary Jean, and Debbie who helped to ease certain difficulties for myself and my family.

Finally, I wish to thank my family for giving me the freedom to complete this degree at a time when I was most needed as a wife and mother. The source of Brian's eternal

patience, I do not know, but his stability made this project a reality. To Sara and Sam I offer my love and soon my time. To Jacob, who taught me perspective, I dedicate this dissertation.

The encouragement, support, and love of my committee, friends, and family are the factors which made the completion of this project possible. Thank you.

TABLE OF CONTENTS

LIST OF TABLES

Abstract of Dissertation Presented to the Graduate School of the University of Florida in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

AN ANALYSIS OF THE PERFORMANCE OF ABOVE AVERAGE READERS AND BELOW AVERAGE READERS ON A PROGRAM-DEPENDENT MASTERY TEST

By

Darla V. McCrea

December 1983

Chairman: William Hedges
Major Department: Instructional Leadership and Support

This study examined the relationship between pupil performance on a program-dependent mastery test in reading and overall reading ability, school assignment and their interaction. This study was designed to determine whether a competency-based program in reading reduces the usual normal distribution of reading achievement when all pupils are allowed varying amounts of time to master program objectives.

The dependent variable used was the Ginn Level 10 mastery test (copyright, 1976). Student scores on the total test and eight subtests were analyzed. Independent variables were students' normed reading ability, measured by the Metropolitan Achievement Test (copyright, 1978), schools to which students were assigned, and their interaction. Four hundred and nine tests were collected from grades three

through six. Subgroups of above average (n=121) and below average (n=65) readers were created for the inferential analyses.

A linear regression model was used to test for significant relationships between the dependent and independent variables. A chi-square analysis tested for a significant difference in the proportion of students from each subgroup who attained mastery criterion. Item analyses (p-values, latent-trait difficulty indices, and goodness-of-fit to the one parameter logistic model) were computed and Pearson product moment correlations were computed between nine variables.

Results of the analyses indicated that MAT was a strong predictor of performance on the mastery test (alpha level .01). School assignment was significant for one subtest (Vocabulary I) and one interaction was significant (Decoding I). A significant difference in the proportions of subgroups achieving mastery favored the above average readers.

The researcher concluded the performance of below average readers was significantly lower than above average readers despite additional time in instruction. Although varying the amount of instructional time was not a sufficient intervention, the interaction observed for one subtest and the significant relationship between school assignment and performance on another suggest that teachers in some

schools have devised interventions which minimize student dependence on overall reading ability.

Research is need to ascertain which factors were successful interventions. Research is also needed to determine how best to evaluate competency-based instructional programs and testing components which accompany them.

CHAPTER ONE

INTRODUCTION

The Problem

This study examined the untested assumption within mastery models of instruction that students' differential aptitudes for learning specific skills--such as reading skills--are irrelevant for mastery of instructional content so long as sufficient time is allowed for learning and the testing to ascertain mastery is content specific. Informal observations of this research have raised doubts about the soundness of this assumption.

The Purpose

The primary purpose was to determine how variation in performance on a program-dependent mastery test is related to overall reading achievement and how this performance is moderated by school assignment. A supplemental investigation examined the items which differed for above average and below average achievers in terms of selected item characteristics. The central question of this study was

> Does the linear combination of the variables reading achievement, school assignment, and their interaction account for a significant proportion

of variance on a program-dependent mastery test for which all students have received instruction? The theoretic rationale for this investigation is discussed in the following section.

Need for the Study

Student achievement in reading is usually assessed through norm-referenced achievement tests. These tests are constructed to yield a substantial distribution of scores in reading achievement for a given age level. Test results for individuals are interpreted through the use of percentile ranks, stanines, or grade equivalents which permit comparison to the total norm group. Because the use of normed achievement tests is extensive, achievement in many content areas such as reading has become accepted to be normally distributed throughout an age group.

On the other hand, mastery testing in reading which accompanies current competency-based reading programs disregards the normal variations in student reading ability. All students who have been instructed are expected to master a certain percentage of skills regardless of overall reading ability. These tests, criterion-referenced in design, may be used in schools to monitor the effectiveness of instructional programs. Usually termed mastery tests, these tests are the most frequently used criterion-referenced tests in practice (Berk, 1978).

Although much research has been done on criterion-referenced tests in general, little attention has been given to the unique qualities of a program-dependent mastery test (Brittain, 1981). The purpose of a program-dependent mastery test is to test only the objectives which accompany a specific instructional program.

The theoretical basis for the design of sequential skill development and mastery testing comes from mastery learning theory (Block, 1971; Bloom, 1976; Bobbitt, 1918; Carroll, 1963; Charters, 1923). This theory has been made operational in competency-based programs for teaching skills such as reading and math. Within a competency-based program, objectives are designed in a hierarchical order. The hierarchy developed for each program is based on sequential skill development and is usually invariant from student to student. That is, students must master skills at the bottom of the hierarchy to be prepared to master skills farther up the hierarchy. This type of planned, invariant hierarchy of skill development is broken into levels for instructional purposes. Mastery at each level is considered necessary for a student to progress to the next level.

To facilitate mastery of each level, the primary objective of the mastery learning model of instruction is to allow each student as much time as necessary to learn and master any skill level before advancing to the next level in the learning hierarchy. An equation to exemplify this

theory of academic achievement was developed by Carroll (1963):

$$\text{degree of learning} = f\left(\frac{\text{time spent to learn}}{\text{time needed to learn}}\right)$$

According to this formula, time is one major difference between the mastery learning model and non-mastery models of instruction.

Supporters of mastery learning theory believe the majority of students can learn a majority of school tasks, regardless of students' individual differences, if enough time on task is allowed to assure mastery (Horton, 1981). Therefore, in theory at least, the effectiveness of the mastery learning model is not limited by the distribution of a given trait, in this case reading achievement, within the norm group. Such distributions should not inhibit the successful mastery of objectives which are taught if sufficient time is given to each child to learn and master an objective.

Research examining the traditional, non-mastery approach to learning supports the theory that under non-mastery conditions in which all students receive the same instructional time, the normal distribution of an academic trait is a significant factor in the achievement of each student. A strong relationship exists between each student's entering aptitude and final performance (Carroll, 1963; Torshen, 1977). Although instruction occurs, if time

spent in instruction is equal for all students, the instruction will not change the distribution of achievement. Those students who enter knowing less will leave knowing less and be less prepared to begin learning at the next level of study.

The mastery learning model of instruction was designed to ensure that all students master objectives taught and that they will be prepared for future instruction. The model is comprised of six components: organizing objectives, preassessment, instruction, diagnostic assessment, prescription, and postassessment (Torshen, 1977). These six components are interrelated and therefore dependent on each other. If the mastery learning model is not effective, that is, if students are not able to show mastery of objectives on postassessment, evaluation of all six components is necessary to determine the cause of failure. Because of this interrelationship between and among components, accuracy in the development and evaluation of the components is crucial. For example, defining objectives and establishing the hierarchy through which students will progress should have a strong theoretical base in that content area as well as in curricular theory. The instructional and prescription components should draw from current knowledge of effective instructional models and learning theory. Finally, the three components which are used for assessment--preassessment, diagnostic assessment, and postassessment--need a proper measurement basis. For these

three components to properly identify student deficiencies or gains, testing materials must be constructed which are accurate and valid for these uses.

The mastery tests which are designed for use in the mastery learning models are criterion-referenced in design and should meet appropriate psychometric requirements for criterion-referenced tests (see Berk, 1980; Hambleton, Swaminathan, Algina, and Coulson, 1978; and Linn, 1979 for reviews of reliability estimation procedures for criterion-referenced and mastery tests). In addition, content validation is required to assure that accurate domain definition and item generation procedures have been used in the construction of the test (Millman, 1974) and the assignment of mastery status on the basis of a test score requires research into the criterion-related and construct validity of the test (Linn, 1979).

Instructional theorists believe two other types of validity are important to program-dependent mastery tests (McClung, 1977). First, the test must possess curricular validity. This is established when the tested objectives can be found in the objectives of the established curriculum being taught. For example, if a reading mastery test is measuring decoding of blends, the content of blends must exist in the reading curriculum. Second, instructional validity of tested items must exist. Not only should blends be part of the outlined objectives, they must be taught.

These two types of validity--curricular and instructional--are what make program-dependent tests unique. Through the established hierarchy of skills, which are in turn tested, curricular validity is established. The mastery learning model components of instruction and prescription assure instructional validity.

The application of existing criterion-referenced research to the construction of classroom level mastery tests has been haphazard. Hambleton and Eignor (1978) found many deficiencies in test validation for criterion-referenced tests used commonly in public schools. Noteworthy were their conclusions that reliability of the tests and validity of test score use are questionably determined and reported. More critical evaluations have come from theorists concerned with reading instruction (Brittain, 1981; Shuy, 1982; Walmsley, 1979). These researchers contend that even if basic psychometric analyses were complete and acceptable in a measurement sense, such tests probably do not test the true process of reading. Walmsley's research on establishing adequately defined domains of reading so that items for testing can be developed indicated that establishing separate domains may not be possible due to the fractionalization this imposes on the reading process. Brittain (1981) reemphasized this point and questioned the notion of invariant hierarchies built into commercial competency-based reading programs. Shuy believes tests presently being used may be inadequate

to draw inferences about student reading ability. He suggested that test results are superseding the judgment of good reading teachers. This overemphasis on test scores may focus attention away from reading content or more extensive teacher training (Shuy, 1982).

The researchers discussed above raise doubts as to the feasibility that mastery testing solely can be used for inferences about student reading ability and for monitoring student reading progress. If the validation of test use is incomplete, and if the ability to test the process is suspect, can the results be used to accurately monitor either a competency-based reading program designed according to the mastery learning model or the individual students within that program?

A need therefore exists to test the basic assumption existing within the mastery learning model as it is represented in current development of program dependent criterion-referenced tests. First, does a competency-based instructional program tend to equalize the normed distribution of the trait of reaching achievement? That is, do students who have been instructed in a specific level of objectives geared to their functional level perform similarly regardless of their overall reading ability? And second, based upon examination of item statistics, what are the structural or content characteristics of items on the mastery tests which function differently for above and below average readers who have received the same instruction?

In this study, an analysis was made of students' total score, subtest scores and responses to individual items on a program-dependent mastery test in reading which accompanied a widely used commercial competency-based reading program. All students had received instruction on the objectives for this test. This reading program fits the mastery learning model in that all six components are presented and implemented. Above average and below average readers were identified on the basis of Metropolitan Achievement Test (MAT) percentile ranks.

Limitations

The sample used in this study was taken from one North Florida county. Therefore, generalizability of results is limited to similar populations. This study analyzed test results from one test, Level 10 of one basal reading series--Ginn and Company, 720 Reading Edition. Although this test fits the description of a program-dependent test, the specific results of these analyses cannot be generalized to other published tests of this type without similar analysis.

Definitions

A criterion-referenced test is one which "depend(s) upon an absolute standard of quality" (Glaser, 1963, p. 519) for interpretation of test scores.

A mastery test is "a criterion-referenced test used to ascertain an individual's status with respect to a well defined behavioral domain" (Popham, 1978, p. 93).

A program-dependent test is a test organized around objectives sequenced in a hierarchical arrangement with items keyed to each objective (Brittain, 1981).

A compentency-based instructional program identifies objectives, sequences these objectives, instructs students, and tests for mastery of objectives. Students instructed using a competency-based instructional program enter the skill hierarchy at their functional level and are not progressed until mastery of each level is achieved.

Above average readers have been identified for this study as students who achieved a percentile rank of seventy-seven or higher on the Meteropolitan Achievement Test in reading.

Below average readers have been identified for this study as students who achieved a percentile rank of thirty or below on the Metropolitan Achievement Test in reading.

School Assignment in this study is used as a control variable and refers to the school which a student attends. Students from nine elementary schools participated in this study.

Mastery status is a variable used in this study to designate whether or not students achieved the mastery level criterion on the Ginn Level 10 mastery test. The criterion used in this study was a minimum score of seventy-three out

of the ninety-one items. This criterion was used because it is the recommended criterion set by Ginn and Company for this level test.

Assumptions

The major assumption of this study is that the prescribed instruction necessary for taking the program-dependent test took place within each classroom prior to testing. The monitoring done in each elementary school by curriculum resource teachers, combined with the mastery learning model component structure of the Ginn 720 Basal Reading Series, assures that instruction did take place. However, the use of every instructional activity in the reading series could not be verified as occurring for all students in the sample.

A second assumption of this study was that a single linear model could be used to characterize the relationship between the dependent variable and the predictor variables of interest.

Organization of the Study

This chapter presented the problem under study and a rationale for investigating this problem. A review of the germane literature is presented in Chapter Two. This review covers three areas: mastery learning theory, criterion-referenced testing, and item analyses of criterion-referenced tests. The research methodology used in this

study is discussed in Chapter Three. This discussion covers the topics of instrumentation, subject selection, and data analyses. A report of the results of these analyses is contained in Chapter Four. Chapter Five contains a discussion of the results, conclusions, and recommendations for future research.

CHAPTER TWO

REVIEW OF THE LITERATURE

Although research has been done on the psychometric properties of criterion-referenced tests, limited attention has been given to the interaction of such tests with curriculum content areas or the impact of criterion-referenced tests on instructional design. The discussion in Chapter One outlined the problem under study. The literature review presented in Chapter Two discusses current research germane to the problem. Three topics are reviewed in Chapter Two: mastery learning theory, general practices in criterion-referenced measurement, and item analyses for criterion-referenced tests.

Mastery Learning Theory

Mastery of program objectives is the goal of competency-based programs (Torshen, 1977, p. 41). Mastery learning theory is an approach to structure curricula and instruction to ensure all students attain acceptable levels of performance in competency-based programs. This model is based on the proposition that a majority of students can learn the basic skills in school curricula when instruction

is of good quality, appropriate, and when adequate time is spent on learning (Torshen, 1977).

Mastery learning theory has evolved from the works of twentieth century educators as Block (1971), Bloom (1976), Bobbitt (1918), Carroll (1963), Charters (1923), and Tyler (1950). The overall emphases of these theorists are that goals need to be established, and objectives, from which instruction can be planned, need to be pinpointed and sequenced. Once instruction is planned, adequate learning time must be provided for learning to occur. Finally, assessment must be given to determine the extent of learning.

The mastery learning model consists of six components: objectives, preassessment, instruction diagnostic assessment, prescription, and postassessment (Torshen, 1977, p. 41). These six components are interdependent. Therefore, for a mastery based program to achieve its goal--acceptable levels of performance in a competency-based program--each of these six components must be evaluated and validated.

Operational difficulties of the mastery model proposition exist due to the implication by theorists that no longer is it acceptable to have large percentages of students who fail to master curricular objectives. Instructional practices such as teaching to the average child, pacing total classes through an established curriculum regardless of ability levels, and lack of adequate

assessment to provide feedback on instructional quality are targets of advocates of the mastery model. If students are to master objectives prior to moving to new content, the factors of time, individual student aptitude, and accurate assessment measures become important considerations.

Time spent in learning is crucial for the success of the mastery learning model. Both the amount of school time devoted to on task instruction as well as the flexibility in the amount of time each student requires have become important aspects to mastery learning theorists. Carroll (1963) defined the degree of student learning as a function of the time spent in learning compared to the time necessary to learn. Through this definition Carroll suggested that almost all students could attain mastery of prescribed objectives if given adequate time. Bloom (1971, 1976) noted that time on task (time spent in active learning) directly relates to the amount a student learns.

In order to plan for flexible amounts of instructional time, the importance of identifying individual aptitudes becomes necessary. Bloom (1971, 1976) lists the acknowledgement of differing cognitive entry levels as a major factor in promoting mastery of basic skills. Bloom argued that if students are normally distributed with respect to a certain trait, given the same instruction and instructional time, then final performance will be normally distributed

as well. That is, a strong relationship would exist between entering aptitude and final performance (Torshen, 1977, p. 50).

By following mastery learning principles however, this relationship between entering levels of aptitude and final performance should diminish. When instructional time is varied to student needs, most students should attain mastery. Research indicates that cognitive entry characteristics do tend to affect postassessment performance (Bloom, 1976; Block & Anderson, 1975). Other research supports the premise that a mastery learning approach can intervene and reduce this relationship between entering aptitude and postassessment performance (Torshen, 1977, p. 72).

The third important factor in operationalizing the mastery learning model is proper assessment to provide information on student aptitude as well as to determine performance on content. Assessments used for this purpose are usually criterion-referenced in design although normative data may be used to determine entry level abilities. A thorough review of criterion-referenced assessments is discussed later in this chapter.

Although the mastery model of instruction has fostered numerous educational programs such as non-graded schools, individualized instruction, and programmed learning packages, criticism of the theory does exist. Most criticism is directed at the limitations such a model might impose on learning. That is, do hierarchical, preplanned

objectives limit thought processes in students? For this reason, critics believe competency-based programs relying on mastery learning principles should be restricted for specific, universally needed basic skills (Anastasi, 1976; Cronbach, 1971). This interpretation of mastery models suggests the mastery learning concept is most effective for subjects which require rote learning and which emphasize convergent thinking.

General Practices in Criterion-Referenced Measurement

Thorndike (1913) initiated the theoretical basis of criterion-referenced measurement. Although sporadically reintroduced during the twentieth century by other psychometricians (Flanagan, 1951; Ebel, 1962) it was Glaser (1963) who operationally defined the difference between norm-referenced and criterion-referenced tests. Glaser distinguished norm-referenced testing from criterion-referenced testing by the standard use as the reference for interpretation of test scores. In norm-referenced measurement the standard is relative. That is, individual score interpretation is dependent on the norm group. In criterion-referenced measurement, the set standard is absolete. In a criterion-referenced system scores are interpreted without the performance of peers being a relevant factor. Rather, individual performance is compared to a set criterion; this criterion is usually set a priori.

Variations in the definition of criterion-referenced measurement exist (Harris & Stewart, 1971; Millman, 1974) because of alternative views of domain definitions and item generation techniques. However, there is general agreement that the distinction between norm-referenced and criterion-referenced testing is "whether the comparison of the score is made to other individuals' scores (norm-referencing) or to some specified standard or set of standards (criterion-referencing)" (Mehrens & Lehmann, 1969, p. 50). The goal of criterion-referenced measurement is not to distinguish among individuals but "to discriminate among [sic] those who have and have not reached set standards" (Mehrens & Lehmann, 1969, p. 51).

Criterion-referenced tests which are used to dichotomously separate examinees into masters/non-masters of specified objectives are called mastery tests. The required level of performance for determining mastery status (cut-off scores) is set a priori. The determination of the placement of the cut-off score is an area of current psychometric research. This cut-off point should not be arbitrary (Glass, 1978). Walmsley (1979) noted that if the usual cut-off for mastery (80 percent correct) is used, in most categories poor readers would be expected to perform better than good readers actually perform. Setting standards is a question of criterion test score validity because these standards affect the accuracy of classifications of individuals. An

extensive discussion of the topic of setting cut-off scores can be found in Berk (1978), Chapter 4 by Hambleton.

Test reliability measures for criterion-referenced tests have been heavily researched. Popham and Husek (1969) pointed out that classical methods for determining reliability would yield lower reliability estimates on criterion-referenced tests because of the dependence of these measures on score variability. Although score variability is not a necessity for a good criterion-referenced test (Linn, 1979) the lack of this variability causes difficulty in determining reliability in traditional ways.

Psychometricians have proposed alternative indices for calculating the reliability of criterion-referenced tests. Livingston (1972) used deviations from the criterion point rather than the mean for calculating reliability. Other theorists argued that reliability of criterion-referenced tests should be viewed in terms of consistency of classification rather than the traditional view of minimal error in measurement. Alternative reliability indices which reflect this viewpoint were developed by Hambleton and Novick (1973) and Swaminathan, Hambleton, and Algina (1974). Huynh (1976) and Subkoviak (1976) developed procedures using single administration of tests. Brennan and Kane (1977) used generalizability theory (Cronbach, Gleser, Nanda, and Rajaratnam, 1972) as a framework for developing a reliability index. Berk (1980) concluded the choice of a

criterion-referenced reliability index depends upon many factors including "test forms assumption, whether or not a cutting score is set . . . intended test score interpretation, type of decision and seriousness of losses associated with the decision errors" (p. 345). Thorough reviews of criterion-referenced reliability indices are in Berk (1980) and Hambleton, Swaminathan, Algina, and Coulson (1978).

Validity for each use of a criterion-referenced test must be established. The uses of test results have been categorized as formative or diagnostic and summative or evaluative (Skager, 1975). Each type of use should be validated separately. In addition, content validation is required to assure that accurate domain definition and item generation procedures have been used in the construction of the test (Millman, 1974). The assignment of mastery status on the basis of a test score requires research into the criterion-related validity and construct validity of the test (Linn, 1979). If mastery of content is necessary as a prerequisite for moving to a new skill, construct validity is needed (Cronbach, 1971). However, if assignment of mastery status implies differential instructional treatments, that is, if the score is used to infer a prediction about this student's potential, criterion validity is needed (Linn, 1979).

Criterion-Referenced Testing in Content Areas

The most extensively used type of criterion-referenced test in practice is the mastery tests (Berk, 1978). These tests accompany published instructional programs and are used at the classroom level for diagnostic purposes or to certify mastery of skills. Mastery tests are criterion-referenced in derivation and therefore should be reviewed for adequacy by criterion-referenced methods.

Hambelton and Eignor (1978) established thirty-nine guidelines pertaining to the evaluation of criterion-referenced tests. They then used these guidelines to review eleven popular criterion-referenced tests in reading and math. They concluded that little evidence was offered for the validity of cut-off scores determining mastery status, reliability was handled inappropriately or not at all, and very few tests discussed error in terms of score stability which affects the consistency of mastery/non-mastery status (p. 325). Hambleton and Eignor also noted that these commercial tests should be labeled "objective-based tests" as described by Popham (1978) rather than criterion-referenced tests because they appear to be developed from behavioral objectives rather than domains. They noted that developing tests from behavioral objectives rather than domains is less than ideal because of the difficulty in establishing item pools (p. 325).

Walmsley did research (1979) on the application of criterion-referenced tests to reading behavior and also

cited problems in the development of item pools. Walmsley described the problem as attempting to fractionalize the process of reading and define numerous, specific universes of reading skills. According to Walmsley, reading theorists are skeptical that fractionalization of reading is theoretically possible. Therefore, what exists is the irreconcilability of a measurement technique--criterion-referenced--which demands that subject matter be precisely defined with the process--reading--which resists efforts to be precisely defined (p. 574). Shuy (1982) discussed this problem of fractionalization in terms of the reading process. According to Shuy, early decoding skills differ from later developmental reading skills (p. 56). Students learn specific sound-letter correspondents, word parts and sentences but later ignore the majority when reading (p. 56). Good readers learn to rely on large cues rather than small cues (p. 57) such as specific decoding skills. Therefore, testing small component skills rather than the totality of reading does not help the teacher (p. 57). Shuy further concluded that reliance on test results for information distracts attention from reading content and teacher training.

In his study, Walmsley explored three important questions. First, can reading universes be defined which fractionalize the reading process for the purpose of testing reading knowledge? Second, is there a way to construct, administer and analyze a test of a specific aspect of

reading? Third, how can results be interpreted from a reading perspective? Walmsley's research led him to conclude that it is possible to define a reading universe in some cases but the definition will "be derived from the test constructor's perspective of the reading process . . . many arbitrary decisions have to be made on what constitutes the dimensions of the universe" (p. 602). In his research, Walmsley tested student knowledge of the structural analysis of CVC (consonant-vowel-consonant) words. Even when narrowing the universe to this small aspect of reading to construct a test, Walmsley encountered difficulty in defining the boundaries of acceptable items. Decisions were made arbitrarily to deal with certain aspects of reading such as nonsense words or the use of CVCC (consonant-vowel-consonant-consonant) words. Walmsley believed such decisions may have significant impact on testing outcomes. Walmsley also noted that constructing a test for structural analysis, even though difficult, was not nearly as complicated as testing comprehension. He listed reading theorists (Drahozal and Hanna, 1978) who emphasized that little empirical evidence is offered to support the test developer's actions of subdividing comprehension into multidimensions. According to these researchers, this lack of evidence makes the results from such tests difficult to interpret in terms of actual student knowledge. Walmsley concluded that a lack of congruence exists between statisical analysis and conceptual perspective of the content matter of reading. He noted that

to this point the statisical procedures used have imposed the conceptual perspective onto the subject matter. That is, the decision to use criterion-referenced measurement to ascertain proficiency of students has led program developers to define content like reading in fractionalized terms in order to make it amenable to testing. Because this fractionalization does not have a theoretical base and may not be appropriate for instruction, score interpretation can be unjustified or at least misleading.

Brittain (1981) supported Walmsley's view that mastery tests in reading fractionalize reading into atomistic parts. She contended that this fractionalization has equated learning to read with mastery of isolated skills. The assignment of mastery or non-mastery status on the basis of the test is questionable in terms of its relationship with reading as a whole.

Brittain noted two other factors pertaining to test content which limit the interpretation which can be made from test results. First, the hierarchical order developed by publishers implies their particular sequencing represents the natural order of learning to read. Such ordering has not been verified. Second, certain subskills are tested whose relationship to reading is not direct. That is, certain skills as color identification on readiness level tests do not necessarily need to be mastered prior to successful reading instruction.

Finally, Brittain noted the paucity of items on any one objective (usually three to five) heightened misclassification of students and that passing scores were set arbitrarily.

Item Analyses for Criterion-Referenced Tests

Logical and empirical item reviews are recommended for all items included on mastery tests (Hambleton, Swaminathan, Algina, and Coulson, 1978). Through the logical review, the relationship between items and objectives is verified. Empirical review is conducted through examination of student responses to items. "The purpose of empirical review is not to select items on the basis of item statistics but to improve items before they are included in the domain" (Haladyna and Roid, 1981b, p. 39).

Research on Item Properties

Item difficulty is an important concern because the level of test difficulty is tied to item difficulties and mastery/non-mastery status is determined by successfully answering certain percentages of questions. Haladyna and Roid (1981a) emphasized that items vary in difficulty as a function of instruction. Therefore, items should be analyzed for their sensitivity to instruction.

Indices have been proposed to assess the instructional sensitivity of criterion-referenced test items. Some of these indices are computed by comparing item responses from

either two testing sessions (pretest and posttest) or two groups (uninstructed and instructed). Cox and Vargas (1966) developed the pre- to post-difference index (PPDI). This index is the percentage difference in pre- to posttest item difficulties. Brennan and Stolurow (1971) suggested using the PPDI to compute a percentage of possible gain (PPG):

$$PPG = \frac{PPDI}{1.00} - \text{pretest difficulty (p value)}$$

Popham (1971) proposed using a phi coefficient to determine instructional sensitivity for each item by using the categories correct and incorrect on pre- and posttests. Haladyna (1974) developed a combined samples point biserial correlation (COMPBI) using instructed and uninstructed students. The size of the coefficient is influenced by the mean difference in total test scores between persons getting the item right or wrong.

Other indices to measure instructional sensitivity have been developed using Bayesian statistics (Helmstadler, 1974) and item response theory (Hambleton and Cook, 1977). A discussion and comparison of these indices can be found in Haladyna and Roid (1981b).

Millman (1978) investigated the relationship between item difficulty and item format and the relationship between item difficulty and language by using computer generated variations of items. He found changes in item content had more effect on difficulty than item format. He concluded

that item difficulties can be made to fluctuate by changes in how questions are asked. More complex questions were more difficult. Millman believed that knowledge about determinants of item difficulty would aid test makers to select those content and format variations to include which would ensure that the skill or domain has been adequately sampled.

Haladyna and Roid (1981a) contrasted the effects of test construction by random sampling of items with items chosen using a latent trait model which matched item difficulty levels to the achievement levels of the students. They hypothesized that random selection of items, which is most frequently used, does not evenly distribute errors of measurement for criterion-referenced tests as is assumed in norm-referenced tests and that the difficulty level of individual items changes with ability levels of students. They also investigated the effect of test length on error of measurement. The results of their research indicated that at-level tests (tests which match student achievement levels with item difficulty levels) consistenly produced the smallest errors of measurement. In addition, test length accounted for a large percentage of the variance and was a powerful factor in reducing error. The function of test length was curvilinear with the greatest decrease in measurement error occurring between ten and twenty items.

Smith (1978) investigated the effects of various item selection methods on classification accuracy and

consistency. By simulating pretest and posttest data on one thousand examinees, Smith varied instructional effectiveness to three levels: students with high ability gaining much and with low ability gaining little, students with no gain, and students with much gain. This study is significant in that it varied the effectiveness of the instruction which could be expected in a normal classroom. Especially pertinent is level one--students with high ability gaining much and with low ability gaining little--since this is probably most reflective of many in-school situations.

Smith used four indices to select forty items on four randomly parallel tests. His results showed that the best item statistic depended upon the level of instructional effectiveness. For level one, varied amounts of instructional effectiveness, the point biserial correlation was best for both accuracy and consistency. Smith's work also showed that variability might exist in actual testing situations to the point where this classical statistic is actually useful. Such variability would be more apt to exist in testing situations which were not specifically program dependent but were assessing general competency levels within a domain.

The issue of variability within the sample is crucial. Shoemaker and Johnson (1981) in assessing construct validity of a district written math criterion-referenced test hypothesized that posttest variance would be greater than pretest variance. This hypothesis was not supported. A possible

explanation is that they did not make the distinction between program-dependent and program-independent tests (Brittain, 1981). Their analysis was of a specific program-dependent test, indicated by their concurrent analysis of time-on task in classrooms spent on the objectives covered by the test. In a program-dependent situation, posttest variance may be reduced since all students are being taught to master the objectives. In addition, Shoemaker and Johnson found that most criterion-referenced test scores correlated with norm-referenced test scores on similar objectives. They interpreted this as evidence for construct validity of their criterion-referenced test. Again, this evidence may be faulty for a program-dependent test. Both tests may be measuring other constructs in similar ways.

Another major focus in item analysis for assessing criterion-referenced test accuracy is to assess the consistency of answers and patterns of responses between items grouped to certain objectives. Although traditional test theory assumes that errors are unsystematic and variability of errors is constant across all examinees (Harnisch, 1981) the examination of item response patterns helps to determine to what extent errors are consistent. Harnisch used two indices: NCI--the consistency between response patterns of an individual and the difficulty ordering for the norm groups, and ICI--the degree of consistency in an individual's response pattern within a topic over time. Harnisch noted that Tatsuoka and Tatsuoka (1980) found that removing

students with low NCIs resulted in a more unidimensional data set. In most cases, students with low NCIs are making either careless mistakes, missing easy questions or are inconsistent in answers. Harnisch found significant differences on NCIs for low ability and high ability students when compared by teacher. This suggested teachers may emphasize different content when teaching similar materials.

Differential item functioning for members of dichotomous groups has been explored through the use of latent trait models (Lord & Novick, 1968). Garcia-Quintana (1981) examined person fit to statewide criterion-referenced assessment data using the RASCH model. Although a very small percentage did not fit the model, the majority of subjects who did not fit had lower than average abilities. Garcia-Quintana concluded that in most cases misfits occurred when low ability students correctly responded to difficult items.

Other studies have investigated differential item functioning by the use of latent trait models. These studies have centered around student characteristics such as race or sex and have relied on normed referenced test data. The purpose of these studies was to identify items which function differentially for members of subgroups. If differential functioning of an item is significant, the item is questionable in terms of validity testing the construct equally for all students. The item should be reviewed for adequacy, revised or thrown out. Various methods have been

proposed to assess differential functioning. Comparisons and reviews of these methods can be found in Shepard, Camilli, and Averill (1981).

Summary

Mastery learning theory suggests that the relationship between the normal distribution of a trait like reading and final performance in a competency based program can be minimized. By assuring adequate learning time and using diagnostic assessments, student mastery of content objectives can be accomplished. Final outcomes are measured through the use of mastery tests, usually criterion-referenced in design.

Research by reading theorists on the use of mastery tests raises doubts about how accurately such tests measure the domain of reading. Two major problems have been pinpointed. First, are currently produced mastery tests constructed and validated adequately? And second, can such tests truly assess the reading process and provide the diagnostic or predictive information which is needed?

Experimental research on criterion-referenced tests has helped to determine empirical methods for establishing test reliability and validity and for reviewing test items. The results of such research indicate that test error can be reduced by using item selection indices which are sensitive to instructional effectiveness and by adequately sampling a domain.

Although a research base exists for developing more dependable criterion-referenced tests for use in competency-based programs, program developers are not incorporating these results into test design. A review of eleven popular criterion referenced tests in reading and math led Hambleton and Eignor (1978) to the conclusion that few acceptable validation procedures were used to assure test quality.

The primary purpose of this study was to test the mastery learning assumption that a competency-based instructional program will minimize the normal distribution of students on a mastery test in reading. A second purpose of this study was to analyze items of a mastery test in reading to identify content and structural characteristics of items on which above average and below average readers differed.

CHAPTER THREE

MATERIALS AND METHODS

The Purpose

The purpose of this study was to determine whether the variation in performance on a program dependent mastery test is related to overall reading ability, school assignment, and the interaction of these variables. Variations in performance were examined at the total test level, the subtest level, and at the item level.

Research Hypothesis

Overall reading ability, school membership, and their interaction account for a significant proportion of the variance in student scores on a program dependent mastery test even when each student receives instruction at his or her developmental level and for varying lengths of time to develop mastery. Factors other than instructional time influence examinee performance.

The following statistical hypotheses were tested as specific components of this general research hypothesis:

1. For below average and above average students in the sample, there is no significant relationship (at alpha

level .05) between total scores obtained on the Ginn Level 10 mastery test and the weighted linear combination of MAT percentile rank, school assignment, and their interaction.

2. For each of the eight Ginn Level 10 mastery subtests, there is no significant relationship (at alpha level .025) between scores on the subtests and the weighted linear combination of MAT percentile rank, school assignment) and their interaction for below-average and above average students in the sample.
3. For those above average and below average readers who take the Ginn Level 10 mastery test in reading, there is no significant difference (at alpha level .01) in the proportions who achieve mastery status.

In addition to the three hypotheses tested, the performance of above average and below average readers was compared on three item parameters for each of the ninety-one items on the Ginn Level 10 mastery test. These parameters included item p-values, item fit to the one parameter logistic latent-trait model, and latent-trait difficulty indices.

Instrumentation

Ginn 720 Basal Reading Series, Level 10 Mastery Test (copyright, 1976)

The program-dependent mastery test used as the major dependent variable in this study was the Level 10 mastery test from the Ginn 720 Basal Reading Series. This test qualified as a program-dependent test because it only measures skill acquisition taught within the hierarchy developed for the Ginn 720 Basal Reading Series. In this particular test, skills taught in Level 10, A Lizard to Start With, are included on the test. This test also qualified as a sample test from a competency-based reading program exemplifying the mastery learning model. All six components of the mastery learning model are incorporated into the instructional cycle of the Ginn reading program.

The test is divided into ten subtests. Eight subtests were used in this study. These eight subtests were chosen because they represent the three major teaching strands of each Ginn reading level. These strands are comprehension, vocabulary, and decoding. The subtests used and the number of test items on each subtest are shown in Table 1.

The Level 10 mastery test was chosen for this study because it spans grade levels in which students have had test-taking practice. Level 10, A Lizard to Start With, is designated a fourth grade level reading book. Due to the mastery learning model, however, students might enter the book as early as third grade or as late as sixth grade.

Table 1

Item Breakdown by Subtest on the Ginn Level 10 Mastery Test

Subject	Skill Tested	Number of Items
Comprehension I	Literal Comprehension	10
Comprehension II	Inferential Comprehension	24
Vocabulary I	Word Meaning	25
Vocabulary II	Context	10
Decoding I	Syllables	5
Decoding II	Digraphs	6
Decoding III	Vowels	6
Decoding IV	Word Parts	5
	Total	91

Entry into Level 10 is determined by attaining mastery of Level 9 skills, or by passing the placement test for Level 10 for students who are new to the school.

No reliability or test validation information is reported in the teacher's test manual. Two contacts with consultants of Ginn and Company have been made. No information on the psychometric evaluation of this test has been made available by Ginn and Company at this time.

Metropolitan Achievement Test (copyright, 1978)

The Metropolitan Achievement Test (MAT) was used to determine reading percentile ranks on each student in the

sample. Level Elementary was administered to third and fourth grade students and Level Intermediate was administered to fifth and sixth grade students. All students used form JS. The reliability estimate, reported in the test manual and based on the Kuder-Richardson Formula 20, is .95 for the Intermediate level and .96 for the Elementary level.

The MAT percentile ranks were used for two purposes. First, the subgroups of above average and below average readers were selected on the basis of percentile rank. Second, student percentile ranks were used subsequently in analyses.

Subjects

The subject population for this study was limited to nine elementary schools in a North Florida county. The administration of each school volunteered to make the data available for the study. The nine schools serve a diverse population including rural and suburban students, low income to professional home backgrounds, and racial and sexual ratios similar to that of the total district. Participants in this study included all students enrolled at each of the nine elementary schools who were given the Level 10 mastery test between April and June of 1982.

Characteristics of the subject population are listed in Table 2.

The percentile ranks from the MAT for all students were used to select the two subgroups of above average and below

Table 2

Grade, Sex, and Racial Composition of Subject Population

	Grade 3		Grade 4		Grade 5		Grade 6		
	Male	Female	Male	Female	Male	Female	Male	Female	Total
Black	4	11	22	34	21	21	5	4	122
White	47	59	94	49	19	11	0	4	283
Other	0	0	1	1	2	0	0	0	4
Total	51	70	117	84	42	32	5	8	409

average readers. In April of 1982, the MAT in reading was administered to each student. This test was a regular part of the sampled school district's student evaluation process. Students with a percentile rank of seventy-seven or higher were chosen for the above average readers. These students fall within the top three stanines for their national norm group. Students with a percentile rank of thirty or lower were chosen as the below average readers. These students fall within the lowest four stanines for their national norm group. Part of stanine four was used for the below average reader group because the mean MAT percentile rank for the total sample was fifty-nine. Therefore, this subject population scored slightly above the national norm group average. Table 3 displays the grade equivalent and scaled

Table 3

Grade Equivalent and Scaled Scores Corresponding to the MAT Percentiles used as Cut-off Points for Below Average Reader and Above Average Readers

	Grade 3	Grade 4	Grade 5	Grade 6
Grade Equivalent	5.3	7.5	9.1	10.1
Scaled Score	707	749	778	797
Above Average Reader Subgroup				
Grade Equivalent	2.7	3.3	4.1	4.5
Scaled Score	621	655	682	693
Below Average Reader Subgroup				

scores corresponding to the percentile ranks used as the cut-off point for each subgroup. Therefore, the grade equivalents and scaled scores displayed are the maximum attained by the below average reader subgroup and the minimum attained by the above average reader subgroup.

By using selection criteria previously described, sixty-five students were identified as below average readers and one hundred twenty-one were identified as above average readers. The composition of these subgroups is described in Table 4.

Table 4

Grade, Sex and Racial Composition of Subgroup Populations

	Grade 3		Grade 4		Grade 5		Grade 6	
	Male	Female	Male	Female	Male	Female	Male	Female
Black	0	0	7	4	9	13	2	3
White	0	0	2	7	13	7	0	1
Other	0	0	1	1	1	0	0	0
Total	0	0	10	12	23	20	2	4
	Below Average Readers (n=65)							
Black	3	7	0	2	0	2	0	0
White	40	40	22	5	0	0	0	0
Other	0	0	0	0	0	0	0	0
Total	43	47	22	7	0	2	0	0
	Above Average Readers (n=121)							

Data Collection

The total sample of 409 Level 10 mastery tests was collected between April and June of 1982. The testing required for this sample collection was part of the instructional cycle within each school. Therefore, no student was required or asked to volunteer for testing. All students tested were students who had been instructed in Level 10 reading objectives and who were recommended for mastery testing by the classroom teacher. Tests were administered, according to the policy of the sampled school district, by the curriculum resource teacher assigned to each elementary school.

Once scoring and recording of scores were completed at the school level, tests were sent to the researcher. A student number was assigned to each test to preserve confidentiality. Each student's grade level, race, and sex were also recorded.

Data Analysis

Descriptive Analysis

Means, standard deviations, and Pearson product moment correlations were computed between total score, the eight subtest scores, and MAT percentile rank using the total sample, the above average readers and below average readers. Correlations were also computed between MAT percentile ranks and Ginn test scores for students in each school subsample.

Hypothesis Analysis

Hypotheses I and II. Hypotheses I and II were tested with multiple regresion using the following linear model:

$$Y = a + b_1X_1 + b_2X_2 + b_3X_1X_2$$

where Y is the dependent variable (score on Ginn test or subtest)

X_1 is the score on the MAT;

X_2 is school assignment; and

X_1X_2 is the interaction between school assignment and score on the MAT.

The values of a, b_1, b_2 and b_3 are respectively the values of the intercept and regression coefficients.

In this regression analysis, above average and below average students were selected from the total student group and the scores of these combined subgroups were entered in the analyses. This was done to increase the power of the analysis without distorting the values of the regression coefficients that would have been estimated from the total group. (See for example, Cramer and Appelbaum, 1978, who note that if a model holds over an entire range of predictor scores, any fixed subset of those predictor scores will yield unbiased estimates of the regression coefficients for the population.) Hypothesis I was tested at the .05 level of significance while hypothesis II was tested at alpha

level .025. A more conservative alpha level was used because of the relatively large number of subtests which were highly correlated. These analyses were computed using the PROC GLM subroutine of the SAS computer program.

Hypothesis III. Hypothesis III tested for a statistically significant difference in the proportion of above and below average readers who attained mastery status on the Level 10 mastery test. A chi-square analysis was computed by the SAS computer program. Hypothesis III was tested at alpha level .01.

Item Analyses

Performance of above average and below average readers was compared on each of the ninety-one items on eight subtests of the Ginn Level 10 mastery test. Item p-values and latent trait item difficulties using the one parameter logistic model were computed separately for each ability group and compared. In addition, each item was analyzed separately by ability group for its fit to the latent trait model. All item analyses were computed using BICAL (Mead, Wright & Bell, 1979).

Review of Items

A structural and content review was made on all items which misfit the latent trait model. The review of structural characteristics included item format and test

directions affecting item responses. The content review focused on linking each item to instructional objectives and activities.

Results of data analyses are presented in Chapter four. A discussion of these results and of the item review is presented in Chapter five.

CHAPTER FOUR

RESULTS AND DISCUSSION

Results of the Analyses

The purpose of this study was to determine the relationship between performance on a program-dependent mastery test in reading for which students had received instruction and the linear combination of the variables of overall reading ability, school assignment, and their interaction. Data were analyzed according to the design outlined in Chapter Three. Descriptive statistics, the results of inferential statistics, and summaries of the item analyses statistics are presented in this chapter.

Results of Descriptive Statistics

Table 5 presents the computed means and standard deviations of the MAT percentile ranks, Ginn Level 10 mastery test scores and subtest scores for the total sample, above average readers and below average readers. The standard deviations, and therefore variances, are greater for the total sample than for the above average readers in all nine test scores presented. This is not true when comparing the total sample calculations with those of the below average readers. In this comparison the standard deviations for six

Table 5

Means and Standard Deviations of the Total Sample (n=409), Above Average Readers (n=121), and Below Average Readers (n=65) for MAT on Ginn Level 10 Tests.

	Total Sample		Above Average		Below Average	
	Mean	S.D.	Mean	S.D.	Mean	S.D.
MAT Percentile	59.2	25.68	90.23	6.4	21.10	8.05
Total Score	75.17	8.5	80.81	5.32	68.41	9.49
Comprehension I	7.09	3.14	7.78	1.51	6.27	1.90
Comprehension II	18.5	4.78	19.83	2.43	16.43	3.36
Vocabulary I	22.83	2.77	24.01	1.36	21.10	3.35
Vocabulary II	8.96	2.42	9.24	1.01	8.13	1.81
Decoding I	4.00	1.00	4.45	.73	3.47	1.17
Decoding II	4.71	1.13	5.23	.89	4.23	1.22
Decoding III	5.25	.93	5.57	.77	4.89	1.00
Decoding IV	4.38	.82	4.70	.47	3.90	1.07

test scores are greater for the below average readers. In all nine tests presented, the computed standard deviations for the test scores of the below average readers are greater than those for the above average readers. This greater variation indicates that among below average readers there was greater variability on the mastery test scores than for above average readers or for the total group.

Table 6 presents the calculated means and standard deviations for all Ginn Level 10 tests by school. Mean MAT percentile ranks for the nine schools ranged from 50.82 to 72.00, indicating considerable spread in average student ability level. Except for school 9 which had only eight subjects, the standard deviations of MAT percentiles ranged from 21 to 31 points. Each school appeared to be fairly heterogenous in terms of student abilities in reading. The mean score on the subtest Comprehension I was below set criterion for all schools in the sample. Seven of the nine schools observed mean scores in subtest Decoding IV which fell below set criterion. School 7 did not reach mastery criterion on five of the nine subtests.

Correlations between Metropolitan Achievement Test percentile ranks, total test score on the Ginn Level 10 mastery test, and subtest scores on the eight subtests of the Ginn Level 10 mastery test were computed using Pearson product moment correlations. Each of the above correlations was computed for three data sets: the total sample (n=409), above average readers (n=121), and below average readers

Table 6

Means and Standard Deviations on MAT, Ginn Level 10 Mastery Test and Subtests by School Membership

School	MAT	Ginn Total	Comp I	Comp II	Voc I	Voc II	Dec I	Dec II	Dec III	Dec IV
1	58.85 (23.17)	75.52 (6.57)	7.20* (2.00)	18.44* (2.33)	22.67 (1.90)	8.88 (.97)	3.76* (1.10)	4.80* (.96)	5.0 (.85)	4.50 (.70)
2	63.32 (25.11)	75.14 (7.20)	6.80* (1.70)	18.05* (2.90)	23.01 (2.10)	9.05 (1.17)	4.09 (.86)	4.75* (1.06)	5.20 (.82)	4.25 (.97)
3	50.82 (27.49)	73.80 (9.90)	6.85* (1.60)	17.33* (3.50)	22.77 (2.60)	8.68 (1.6)	4.17 (1.00)	4.56* (1.20)	5.16 (1.00)	4.30 (.86)
4	51.26 (19.49)	75.12 (7.35)	6.80* (1.70)	18.52* (2.96)	22.80 (1.87)	8.80 (1.08)	4.00 (.96)	4.52* (1.16)	5.16 (.88)	4.50 (.64)
5	69.23 (25.15)	77.36 (6.77)	6.94* (1.82)	19.10 (2.72)	23.46 (1.60)	8.95 (1.26)	4.04 (.93)	4.86* (1.01)	5.37 (.94)	4.59 (.52)
6	63.14 (31.65)	74.14 (9.14)	7.52* (1.69)	18.95* (2.72)	21.38 (3.07)	8.38 (1.35)	3.85* (1.06)	4.66* (1.19)	5.19 (.92)	4.14 (1.15)
7	52.90 (21.36)	72.80* (8.67)	6.95 (1.61)	17.67* (3.07)	21.12 (3.32)	8.80 (1.48)	3.75* (1.23)	4.57* (1.35)	5.32 (1.09)	4.35 (.80)
8	65.00 (28.00)	77.81 (7.91)	7.33* (1.64)	19.14 (2.98)	23.44 (2.22)	8.96 (1.09)	3.96* (1.01)	5.00 (1.07)	5.60 (.62)	4.44 (.69)
9	72.00 (14.02)	79.37 (2.87)	6.75* (1.48)	20.37 (2.13)	23.87 (.99)	9.12 (1.12)	4.37 (.74)	5.00 (.92)	5.25 (.70)	4.62 (.74)

*Mean score for school fell below suggested mastery criterion

(n=65). Tables 7, 8, and 9 present the results of these correlations. Correlations between Ginn test scores and MAT percentile ranks were also computed for each school in the sample.

Numerous correlations were significant at alpha level .01. Out of forty-five possible correlations, thirty-two were significant in the total sample data set. Thus, many of the GINN subtests are significantly related to each other, and to total score. The two variables with the most frequent significant correlations were total score and MAT percentile rank. Twenty-three correlations were significant for the above average readers. Whereas total score continued to be a variable most significantly correlated with the other nine variables, MAT percentile rank was replaced by Comprehension II (inferential comprehension) and Vocabulary II. Finally, eighteen correlations were significant in the data set for the below average readers. Here, MAT percentile rank failed to correlate significantly with any Ginn total or subtest scores although a number of the Ginn subtest scores were highly correlated to each other.

Table 10 presents the correlations between all Ginn test scores and MAT percentile ranks by school. Three negative correlations were observed although none of these was significant. Two of the negative correlations occurred for the same school. This school had a very low number of observations; this could have caused fluctuations resulting

Table 7

Correlations for Total Sample (n=409) Between Total Score on the Ginn Level 10 Mastery Test, Scores on Eight Subtests of the Ginn Level 10 Mastery Test, and MAT Percentile Ranks

	MAT%	TOTSC	Comp I	Comp II	Voc I	Voc II	Dec I	Dec II	Dec III	Dec IV
MAT Percentile	1.00	.537*	.161*	.249*	.339*	.126*	.366*	.302*	.268*	.350*
Total Score		1.00	.046	.216*	.454*	.058	.600*	.539*	.497*	.543*
Comprehension I			1.00	.783*	.543*	.792*	.014	.073	-.023	-.068
Comprehension II				1.00	.589*	.772*	.072	.153*	.007	.065
Vocabulary I					1.00	.639*	.330*	.234*	.177*	.217*
Vocabulary II						1.00	.102	.075	-.005	-.028
Decoding I							1.00	.329*	.308*	.287*
Decoding II								1.00	.266*	.247*
Decoding III									1.00	.244*
Decoding IV										1.00

*Indicates a significant correlation at alpha level .01.

Table 8

Correlations for Above Average Readers (n=121) Between Total Test Score on the Ginn Level 10 Mastery Test, Scores on Eight Subtests of the Ginn Level 10 Mastery Test, and MAT Percentile Ranks

	MAT%	TOTSC	Comp I	Comp II	Voc I	Voc II	Dec I	Dec II	Dec III	Dec IV
MAT Percentile	1.00	.383*	.308*	.332*	.157	.186	.150	.142	.098	.214
Total Score		1.00	.566*	.733*	.551*	.541*	.521*	.371*	.396*	.358*
Comprehension I			1.00	.319*	.053*	.268*	.201	.142	.176	.164
Comprehension II				1.00	.233*	.268*	.244*	-.002*	.163	.306*
Vocabulary I					1.00	.274*	.268*	.160	.133	.250*
Vocabulary II						1.00	.297	.237	.041	.033
Decoding I							1.00	.278*	.142	.080
Decoding II								1.00	.102	-.007
Decoding III									1.00	.147
Decoding IV										1.00

*Indicates a significant correlation at alpha level .01.

Table 9

Correlations for Below Average Readers (n=65) Between Total Test Score on the Ginn Level 10 Mastery Test, Eight Subtest Scores on the Ginn Level 10 Mastery Test, and MAT Percentile Ranks

	MAT%	TOTSC	Comp I	Comp II	Voc I	Voc II	Dec I	Dec II	Dec III	Dec IV
MAT Percentile	1.00	.269	.183	.168	.154	.281	.128	.140	.287	.064
Total Score		1.00	.678*	.708*	.747*	.683*	.485*	.442*	.492*	.495*
Comprehension I			1.00	.494*	.309	.413*	.212	.340*	.310	.127
Comprehension II				1.00	.239	.339*	.073	.249	.282	.349*
Vocabulary I					1.00	.521*	.426*	.188	.319*	.406*
Vocabulary II						1.00	.372*	.112	.223	.224
Decoding I							1.00	.237	.256	.184
Decoding II								1.00	.212	.123
Decoding III									1.00	.136
Decoding IV										1.00

*Indicates a significant correlation at alpha level .01.

Table 10

Pearson Product Moment Correlations of MAT Percentile Ranks with Ginn Test Scores by School

School	n	Total Score	Comp I	Comp II	Voc I	Voc II	Dec I	Dec II	Dec III	Dec IV
1	32	.445*	.324	.320	.416	.331	.167	.039	-.007	.306
2	77	.596*	.421*	.505*	.423*	.199	.206	.208	.334*	.404*
3	80	.545*	.385*	.445*	.397*	.305*	.508*	.345*	.324*	.327*
4	50	.413	.189	.311	.194	.346	.453*	.153	.360	.176
5	69	.555*	.367*	.377*	.401*	.299	.273	.369*	.164	.481*
6	21	.721*	.393	.475	.648*	.414	.662*	.374	.380	.581*
7	40	.497*	.388	.259	.379	.325	.453*	.309	.117	.070
8	28	.594*	.340	.493*	.448	.125	.474	.409	.390	.425
9	8	.779*	-.390	.849*	.452	.018	.766	.220	-.634	.301

*Indicates significance at alpha level .01.

in the negative correlations. Thirty-three of the eighty-one possible correlations were significant at alpha level .01. Of the nine schools in the sample, eight had significant correlations between total test score and MAT percentile ranks. The subtest with the most frequent significant correlations was Comprehension II (inferential comprehension).

Findings Related to the Hypotheses

Hypotheses I and II were tested with multiple regression using the following linear model:

$$Y = a + b_1X_1 + b_2X_2 + b_3X_1X_2$$

where Y is the dependent variable;

X_1 is the score on the MAT:

X_2 is school assignment; and

X_2X_2 is the interaction between school assignment and score on the MAT.

The values of a, b_1 and b_2 are respectively the values of the intercept and regression coefficients.

The linear model was used repeatedly for the Ginn total score and subtest scores in accordance with Hypothesis I and II using only the restricted sample comprised of the above

average and below average readers. All analyses for Hypotheses I and II were computed using the PROC GLM subroutine of the SAS computer program. Type I sums of squares were used for a hierarchical interpretation of the effects of the model. The results presented in Tables 11 and 12 are the results of testing this multiple regression model using only this restricted sample.

By using this model, the data can be interpreted in three ways. First, is performance on the mastery test significantly related to overall reading ability? Second, is performance on the mastery test significantly affected by school after controlling for differences in overall reading ability? And third, is performance on the mastery test significantly affected by the interaction of school assignment and overall reading ability?

Hypothesis I

Hypothesis I was tested to determine if there was a significant relationship between scores achieved on the Ginn Level 10 mastery test (total score) and overall reading ability, school assignment, and their interaction. The hypothesis was tested at the .01 level of significance.

The overall linear model produced an F statistic of 10.07, $p = .0001$. The percent of variance accounted for was $R^2 = .504$. The interaction and school assignment effects did not contribute significantly to variance in the dependent variable. Overall reading ability was significant

($F = 157.33$; $p = .0001$). The researcher concluded only the variable of overall reading ability was significantly related to scores achieved on the Ginn Level 10 mastery tests at the total test level. Table 11 summarizes the results of testing Hypothesis I.

Hypothesis II

Hypothesis II tested for a significant relationship between scores achieved on the eight subtests of the Ginn Level 10 mastery test and overall reading ability, school assignment and the interaction of these two variables. Hypothesis II was tested at the .025 level of significance. Results of this hypothesis were presented in Table 12.

Table 11

Results of Regression Analysis for Hypothesis I Examining the Relationship of MAT Percentile Rank, School Assignment, and Their Interaction to Performance on the Ginn Level 10 Mastery Test (n=186).

Overall Test of Model			Tests for Components in Model				
R^2	F	p	Variable	Sums of Squares	df	F	p
.504	10.07	.0001*	MAT	7285.43	1	157.33	.0001*
			School	288.38	8	.78	.6222
			Inter.	351.21	8	.95	.4786

*Indicates significance at alpha level .05.

Table 12

Results of Regression Analysis for Hypothesis II Examining the Relationship of MAT Percentile Rank, School Assignment, and Their Interaction to Performance on Subtests of the Ginn Level 10 Mastery Test (n=186)

Dependent Variable	Overall Test of Model R^2	F	p	Tests for Components in Model Variable	Sums of Squares	df	F	p
Comp I	.229	2.94	.0002*	MAT	116.11	1	41.79	.0001*
				School	17.06	8	.77	.6316
				Inter	5.58	8	.25	.9799
Comp II	.36	5.62	.0001*	MAT	574.47	1	77.95	.0001*
				School	84.58	8	1.43	.1853*
				Inter	45.47	8	.77	.0558
Voc I	.41	7.00	.0001*	MAT	378.13	1	83.24	.0001*
				School	91.87	8	1.43	.0127
				Inter	70.84	8	1.95	.6286
Voc II	.20	4.62	.0013*	MAT	63.14	1	34.70	.0001*
				School	8.64	8	.59	.7820
				Inter	6.87	8	.47	.8743
Dec I	.32	4.84	.0001*	MAT	43.19	1	56.12	.0001*
				School	4.92	8	.80	.6034
				Inter	15.22	8	2.47	.0147*

Continued

Table 12

Continued

Dependent Variable	Overall Test of Model R^2	F	p	Variable	Tests for Components in Model Sums of Squares	df	F	p
Dec II	.26	3.55	.0001*	MAT	46.57	1	45.34	.0001*
				School	5.35	8	.65	.7337
				Inter	10.10	8	1.23	.2846
Dec III	.20	2.51	.0015*	MAT	22.69	1	30.79	.0001*
				School	4.11	8	.70	.6925
				Inter	4.68	8	.79	.6078
Dec IV	.31	4.63	.0001*	MAT	28.14	1	54.48	.0001*
				School	6.04	8	1.46	.1745
				Inter	6.48	8	1.57	.1371

* indicates significance at alpha level .025.

Of the eight subtests tested in Hypothesis II, one had a significant interaction. This was the subtest of Decoding I. Decoding I tested the ability to decode four and five syllable words. This significant interaction indicates that performance on this subtest is differentially affected by school membership for certain levels of overall reading ability. To ascertain the nature of this interaction, regression coefficients were used to plot performance by MAT levels within each of the nine elementary schools in the sample. Figure 1 graphs this relationship and is presented in Chapter Five.

The variable of school assignment related significantly to performance on one subtest, Vocabulary I. For this subtest, the school to which students were assigned was a significant factor in performance on the test.

The variable of overall reading ability was significantly related to performance on all eight subtests. The R^2 values ranged from .20 to .41.

Hypothesis III

Hypothesis III was tested to determine if there was a significant difference in the proportion of students identified as above average and below average readers who took the Ginn Level 10 mastery test and achieved mastery status. Alpha level .01 was used.

A chi-square analysis was used to test hypothesis III. Mastery status was divided into master/non-master of Ginn

Level 10. Mastery status was determined by applying the Ginn criterion of answering correctly 80 percent or more of the ninety-one items attempted. By using this criterion, all students achieving a score of seventy-two or below were considered non-masters. Scores of seventy-three or above were classified as masters.

The computed chi-square statistic equaled 59.00, $p = .0001$. Since the probability of obtaining the computed statistic was less than the .01 level set as criterion for statistical significance, the null hypothesis was rejected. There is a statistically significant difference in the proportion of above average and below average readers who achieve mastery status on the Ginn Level 10 mastery test. As might be expected, this proportional difference favors the above average readers. Table 13 summarizes the results of this analysis.

Item Analysis

Three item parameters were used to compare performance of above average and below average readers on the ninety-one individual items on the Ginn Level 10 mastery test: item p-values (the proportion answering the item correctly), latent-trait difficulty indices, and goodness-of-fit to the one parameter logistic item response model. These item parameters were estimated separately for the subgroups of above average and below average readers using the BICAL

Table 13

Results of Hypothesis III: Chi-Square Analysis of Proportions of Above Average Readers and Below Average Readers who Achieve Mastery of Ginn Level 10

Ability Groups	Ginn Level 10 Mastery Status	
	Master	Non-Master
Above Average Readers	60% (n=111)	5% (n=10)
Below Average Readers	13.5% (n=25)	21.5% (n=39)

Computed $X^2 = 59.64$, $p = .0001$

x2*
1, .01 = 6.35

computer program (Mead, Wright, and Bell, 1979). Table 14 contains these results.

Item p-Values

Across all subtests the p-values for items on the Ginn Level 10 mastery test ranged from a low of .46 to a high of 1.00 for students in the above average reader group. This indicates the most difficult item was passed by 46 percent of this ability group whereas the easiest item was passed by 100 percent of this group. Seventy-three of the ninety-one items had p-values of .80 or higher. Therefore at least 80 percent of the above average readers correctly answered 80 percent of the items.

The item difficulties were very different for the below average readers. The p-values of items for this group of students ranged from a low of .33 to a high of 1.00. Only thirty-three of the ninety-one items had p-values of .80 or higher.

With the exception of four items, the p-values of items for the above average readers were all higher than those of the below average readers. Of the four items, two had values which were equal for these groups and two had p-values which were higher for the below average readers.

Table 15 presents the range of p-values by subtest for the above average and below average readers. This table

Table 14

Results of Item Analyses: Item p-Values, Latent Trait Difficulty Indices, and Fit to Latent Trait Model

Subtest	Item	Item p-Values		Difficulty Indices		Fit to One Parameter Logistic Model	
		AAR*	BAR**	AAR*	BAR**	AAR*	BAR**
Literal Comprehension	1	.92	.70	-0.050	.363	- .20	-2.17
	2	.77	.70	1.306	.363	-1.96	2.63
	3	.99	.72	-1.933	.281	1.58	-0.91
	4	.89	.67	.275	.521	.11	-0.58
	5	.57	.44	2.343	1.563	0.68	-1.04
	6	.54	.40	2.457	1.769	1.48	0.41
	7	.89	.76	.369	.014	-0.24	-0.55
	8	.60	.58	2.188	.955	0.46	2.49
	9	.89	.70	.275	.363	0.18	0.18
	10	.70	.49	1.691	1.361	-0.16	-0.01
Inferential Comprehension	1	.89	.67	.275	.521	0.93	-0.53
	2	.71	.64	1.646	.671	0.04	-0.26
	3	.75	.75	1.408	.107	3.03	1.73
	4	.64	.55	1.989	1.091	-0.31	0.87
	5	.80	.76	1.085	.014	0.25	0.15
	6	.83	.70	.901	.363	1.47	1.22
	7	.93	.70	-0.178	.363	-1.98	2.44
	8	.97	.80	-1.137	- .184	0.94	3.80
	9	.87	.66	.540	.596	-0.23	-0.72
	10	.94	.72	-0.480	.281	-0.91	-2.12
	11	.84	.67	.766	.521	2.46	-1.54
	12	.83	.89	.901	- .979	1.51	2.24

Continued

Table 14

Continued

Subtest	Item	Item p-Values		Difficulty Indices		Fit to One Parameter Logistic Model	
		AAR*	BAR**	AAR*	BAR**	AAR*	BAR**
	13	.86	.67	.619	.521	1.15	-1.17
	14	.80	.64	1.085	.671	0.51	0.11
	15	.95	.63	- .663	.743	-0.71	1.11
	16	.80	.70	1.085	.363	2.37	0.43
	17	.75	.56	1.408	1.023	0.36	1.33
	18	.46	.44	2.835	1.563	0.22	2.26
	19	.91	.70	.067	.363	-0.34	-0.21
	20	.76	.49	1.358	1.361	-0.38	-1.02
	21	.69	.33	1.735	2.058	-0.39	0.69
	22	.77	.78	1.306	- .082	0.53	0.09
	23	.98	.72	-1.469	.281	0.04	-0.54
	24	.93	.75	.178	.107	-0.23	-0.20
Vocabulary I	1	.98	.93	-1.469	-3.208	-0.72	-1.28
	2	.98	.93	-1.469	-1.644	1.03	-0.87
	3	.98	.86	-1.469	- .666	0.04	1.25
	4	.99	.87	-1.933	- .814	-0.54	0.48
	5	.94	.87	- .480	- .814	0.54	-0.91
	6	.87	.66	.540	.596	0.43	-0.04
	7	.98	.98	-1.469	-3.208	-0.40	-1.28
	8	.88	.64	.457	.671	-0.79	0.77
	9	.93	.75	- .178	.107	-0.30	0.48
	10	.95	.78	- .663	- .082	-0.22	-0.01
	11	.99	.96	-1.933	-2.435	-0.54	1.20
	12	1.00	.86	-2.719	-0.666	-1.32	-0.80
	13	.96	.84	-0.877	-0.532	-1.02	-0.59
	14	.97	.95	-1.137	-1.975	-0.44	0.18
	15	.96	.73	-0.877	.195	0.01	0.86

Continued

Table 14

Continued

Subtest	Item	Item p-Values		Difficulty Indices		Fit to One Parameter Logistic Model	
		AAR*	BAR**	AAR*	BAR**	AAR*	BAR**
	16	.96	.90	-0.877	-1.165	0.27	0.23
	17	.93	.84	-0.178	-0.532	-0.41	-0.60
	18	.81	.41	1.026	1.700	0.32	-1.58
	19	.98	.86	-1.469	- .666	-1.01	0.78
	20	.99	.84	-1.933	- .532	0.73	-0.05
	21	.99	.90	-1.933	-1.165	-0.54	-1.24
	22	.89	.60	.275	.885	0.15	0.10
	23	.97	.84	-1.137	- .532	0.49	0.57
	24	.98	.93	-1.469	-1.644	-1.06	-0.02
	25	.94	.67	-0.320	.521	-0.28	-0.47
Vocabulary II	1	.85	.61	.694	.815	0.09	-0.63
	2	.95	.96	- .663	-2.435	0.37	0.89
	3	.90	.84	.175	- .532	-0.49	-0.27
	4	.72	.58	1.600	.955	-0.11	0.05
	5	.96	.95	-0.877	-1.975	-0.23	0.21
	6	.97	.87	-1.137	- .814	0.53	0.26
	7	.90	.73	.175	.195	-0.52	0.30
	8	.96	.73	- .877	.195	-1.69	1.08
	9	.94	.78	- .320	- .082	1.11	-0.61
	10	.98	.90	-1.469	-1.165	0.24	0.22
Decoding I Syllables	1	.97	.81	-1.137	- .293	-0.42	0.67
	2	.85	.50	.694	1.294	0.48	1.05
	3	.89	.61	.275	.815	1.19	0.10
	4	.69	.46	1.735	1.495	1.44	-2.25
	5	.93	.90	- .178	-1.165	1.01	0.32

Continued

Table 14

Continued

Subtest	Item	Item p-Values		Difficulty Indices		Fit toOne Parameter Logistic Model	
		AAR*	BAR**	AAR*	BAR**	AAR*	BAR**
Decoding 2	1	.92	.60	- .050	.885	0.96	-1.85
Diagraphs	2	.90	.81	.175	-.293	-0.38	-0.74
	3	.86	.75	.619	.107	-0.07	0.90
	4	.78	.63	1.253	.743	2.23	1.39
	5	.73	.56	1.554	1.023	-0.60	0.89
	6	.92	.72	-0.050	.281	-0.82	-0.56
Decoding 3	1	.97	.98	-1.137	-3.208	-0.60	1.38
Vowels	2	.85	.66	.694	.596	1.53	-1.72
	3	.84	.63	.835	.743	-0.15	2.05
	4	.96	.81	- .877	- .293	-1.02	0.01
	5	.99	1.00	-1.933	none	-1.01	none
	6	.86	.69	.619	.443	0.42	0.17
Decoding 4	1	.97	.83	-1.137	- .408	0.53	-1.09
Word Parts	2	.96	.86	- .877	- .666	0.96	-0.95
	3	.95	.72	- .663	.281	-0.54	0.55
	4	.72	.44	1.600	1.563	1.24	0.11
	5	.98	.92	-1.469	-1.382	-0.72	0.47

*Above average readers.
**Below average readers.

Table 15

Item p-Values on the Eight Subtests of the Ginn Level 10 Mastery Test for Above Average and Below Average Readers

Subtest	Item p-Values							
	Above Average Readers				Below Average Readers			
	Lowest	Highest	Range	Median	Lowest	Highest	Range	Median
Comprehension I	.54	.99	45	.83	.40	.76	36	.69
Comprehension II	.46	.98	52	.83	.33	.89	56	.67
Vocabulary I	.81	1.00	19	.97	.41	.98	57	.84
Vocabulary II	.72	.95	23	.95	.58	.96	38	.81
Decoding I	.69	.97	28	.89	.46	.90	44	.61
Decoding II	.73	.92	19	.88	.56	.81	25	.68
Decoding III	.84	.99	15	.91	.63	1.00	37	.75
Decoding IV	.72	.98	26	.96	.44	.92	48	.83

also presents the median p-value for each subtest, for each subgroup.

Fit to the One Parameter Logistic Model

As a further exploratory procedure to provide insight into the particular items in which above average and below average readers differed, results of an item analysis based on the RASCH one parameter logistic model were examined. The statistic chosen for examination was the goodness-of-fit test which is used to identify items which display a significant degree of misfit to the model. A separate item analysis was run for the above average and below average readers. If the goodness-of-fit statistic exceeded 2.00, the item was identified as a misfit. Results were reported in Table 14. For the above average reader group, four items were found which failed to meet the goodness-of-fit criterion. For the below average readers, ten items failed to fit the model. Results of this analysis do not provide a basis for concluding that the items on the mastery test appear to measure different traits for high and low ability students.

Latent-Trait Difficulty Indices

The difficult indices calculated for each item by subgroup membership are presented in Table 14. A latent-trait difficulty index uses a transformed distribution of scores along a continuum. The computed index for each item

represents the point on this continuum where the probability exists that 50 percent of the examinees would answer the item correctly. Therefore, a negative index represents a lower difficulty level; positive indices represent more difficult items since the probability of answering these items correctly requires greater amounts of the latent trait. Although it might seem that above average and below average readers would tend to have differences in their item difficulties, if the test is measuring the same trait for all students these difficulty indices which are generated by the latent-trait model should be closely associated. In fact, if all items are perfect fits to the model and measure the same trait for both groups, the latent-trait difficulty estimates for the two subgroups should differ only by a constant amount. Thus the estimated difficulty indices for the above average readers would be a simple linear transformation of the difficulty estimates for the below average readers.

In this study, once the difficulty indices were calculated for each group, the researcher correlated the item indices for the total test (n=91) and again for the total test minus the fourteen items which misfit the latent-trait model (n=77). The correlation of item indices for all ninety-one items was .735 and for the seventy-seven items was .744. Although there are no standard guidelines for interpreting these correlations these results are probably

too high to warrant a conclusion that these items measure different traits for these two subgroups.

Summary of Results

In summary, findings of this study indicate that performance on the Ginn Level 10 mastery test is significantly related to overall reading ability. This conclusion is supported at the total test level (Hypothesis I), subtest level (Hypothesis II), and through item analyses. In addition, above average and below average readers differ significantly in the proportion of examinees who achieve mastery level on level 10. Evidence exists (Hypothesis II) that the relationship between overall ability and performance on this mastery test can be affected by school assignment. This interaction suggests that this variable may change the relationship between aptitude and achievement for this program-dependent mastery test. Factors within the variable of school assignment which may be the cause of this interaction can only be speculated. A discussion of these results is presented in Chapter Five.

CHAPTER FIVE

DISCUSSION, CONCLUSIONS, AND RECOMMENDATIONS

The primary purpose of this study was to determine to what extent pupil performance on a program-dependent mastery test is determined by their overall reading ability, school assignment, and the interaction of these two variables. The amount of instruction each child received varied and depended upon the individual needs of the child. The problem which prompted this study was the untested assumption in mastery learning programs of instruction that varied amounts of instruction, if given at each student's developmental level, will be a sufficient intervention for students with lower aptitudes.

Three hypotheses were tested in this study. The results of these analyses were presented in Chapter Four. This chapter will present a discussion of the results, conclusions about the problem under study, and recommendations for future research and pedagogical practice.

Discussion of Descriptive Analysis

Pearson product moment correlations were computed between ten variables for the total sample (n=409), for the above average readers (n=121) and for the below average

readers (n=65). These results are presented in Tables 7, 8, and 9.

For the total sample, Metropolitan Achievement Test percentile ranks were significantly correlated with the total test score and all subtest scores of the mastery test examined in this study. The high frequency as well as strength of these correlations suggests that within this sample population, higher scores on the mastery test were associated with greater reading ability. This frequency of significant correlations with the MAT percentile ranks was not true for the two subgroups of above average and below average readers. The lack of correlations for these groups may be due to the restricted range of scores in each subgroup.

Although two of the eight subtest scores were not significantly correlated with total score for the total sample, all eight subtests were significantly correlated with the total score for the above average and below average readers. Comprehension II (inferential comprehension) and Vocabulary II (context) were the two subtests most frequently correlated with the others at levels of significance (alpha .01). These correlations suggest that all subtests and items in this mastery test, regardless of the intended skill to be tested, may depend upon contextual reading abilities as well as the ability to make inferences. That is, the overall ability to infer and read contextually may be necessary to accomplish other reading tasks well.

Pearson product moment correlations were computed between MAT percentile ranks and each Ginn test score for each school in the sample (see Table 10). For school 3, all subtests were significantly correlated with MAT percentile rank. This result suggests that at this school, overall reading ability was a strong determinant of performance on all Ginn Level 10 subtests. The subtest of Comprehension II was significantly correlated with MAT percentile ranks for five schools in the sample and was the subtest which correlated the most frequently at significant levels.

One school in the sample, school 4, had only one subtest that correlated significantly with MAT percentile rank. This result suggests that overall ability levels within this school were not strong determinants of performance on the Ginn Level 10 mastery test. Factors other than reading ability account for reading performance in school 4; these other factors may be reducing the dependence upon reading ability to perform well on the program-dependent mastery test.

Table 5 presents means and standard deviations for all nine tests, for all three samples. The standard deviations for these three groups should be noted. The largest standard deviations should be expected from the total sample because it is larger and more heterogenous and smaller standard deviations should be expected for the more homogenous samples of above average and below average readers. This expectation was true when comparing the

standard deviations of the above average readers with those of the total sample. The standard deviation for each test for the above average readers is smaller than the deviations for the total sample. This reduction in variance did not occur consistently for the below average readers. The standard deviations for the below average readers were always larger than those of the above average readers and larger in six of nine scores than the total sample. Although this sample was smaller, and spanned a smaller range of scores, a much greater variance in scores was observed. This greater variance may be the result of the occurrence that some below average readers are attaining mastery scores even though many are not. That is, the below average readers are not consistently, as a group, failing or succeeding on any one subtest or on the total test.

Tables 16 helps to interpret the importance of the means attained by ability groups on each subtest and on the total test. The means on a program-dependent mastery test become very important when compared to the criterion set as mastery for each test. Since a program-dependent test does not show general levels of ability but instead are interpreted in terms of mastery of a specific skill, the means achieved by a group indicate whether that group, as a whole, has attained mastery of the required skills. Table 16 lists the score set as criterion for each test. The criterion used for each test is the criterion suggested by Ginn and Company for this test and is based on an 80 percent mastery

Table 16

Comparison of Means of Total Sample (n=409), Above Average Readers (n=121), and Below Average Readers (n=65) to Suggested Mastery Criteria on the Ginn Level 10 Mastery Test and Subtests

Test	Number of of Items	Criteria for Mastery	Total Sample		Above Average		Below Average	
			Mean Score	Difference	Mean Score	Difference	Mean Score	Difference
Total Test	91	73	75.17	+2.17	80.81	7.81	68.41	-4.5*
Literal Comprehension	10	8	7.09	- .91*	7.78	- .22*	6.27	-1.7*
Inferential Comprehension	24	19	18.5	- .50*	19.83	.83	16.43	-2.5*
Vocabulary Word Meaning	25	20	22.83	2.83	24.01	4.01	21.10	1.1*
Vocabulary II Context	10	8	8.96	.96	9.24	1.24	8.13	.1
Decoding I Syllables	5	4	4.00	0.00	4.45	.45	3.47	- .5*
Decoding II Digraphs	6	5	4.71	- .29*	5.23	.23	4.23	- .7*
Decoding III Vowels	6	5	5.25	.25	5.57	.57	4.89	- .1*
Decoding IV Word-parts	5	4	4.38	.38	4.70	.70	3.90	- .1*

*Means which did not meet mastery criteria

rate. The table then compares these set criteria to the means attained by the total sample, the above average readers and the below average readers.

The most important observation in Table 16 is that no group attained mastery level means on every subtest. The total group did not attain mastery level means on three of the tests, the above average readers missed criterion on one test, and the below average readers missed mastery criterion on each of seven tests. No group attained a mastery level mean for the subtest of Literal Comprehension. These results tend to support Walmsley's argument (1979) that poor readers are expected to perform better than good readers actually do perform.

Table 6 presented the means and standard deviations on all tests by school. Again, no mean achieved by schools on the subtest of Literal Comprehension met the set criterion. The set criterion was not met for Decoding II by students in seven of the nine schools in the sample. Students in one school did not achieve a mean on the total score which met the set criterion.

The difficulty that students in all groups and all schools had on the subtest of Literal Comprehension suggests either the entire sample was not well prepared for the objectives tested in this subtest, the test items did not accurately test the objectives for which the students had prepared, or these objectives are too difficult for this population to achieve. Since the above average readers, and

the majority of school means exceeded mastery criteria on other subtests it seems unlikely that students were not prepared on this material. Thus it would seem that the subtest of Literal Comprehension may not be accurately testing the objectives being taught or these objectives are not appropriate. Further research is necessary to ascertain which of these hypotheses is correct.

The below average readers attained means which reached mastery criterion on only two subtests, both vocabulary. For the other seven tests, the group mean did not reach criterion. Therefore, for this sample of below average readers, the relationship between overall reading ability, as measured by the Metropolitan Achievement Test, and final performance in the competency-based reading program, as measured by the Ginn Level 10 mastery test was not diminished. Rather, students with below average reading abilities tended to score low and not achieve mastery levels on most areas of the Ginn Level 10 mastery test, despite having progressed through the instructional program.

Summary

Results of descriptive analyses support two main conclusions. For this sample of students, general reading ability was strongly correlated to mastery of Ginn Level 10 objectives. Students with less reading ability had difficulty reaching mastery criteria regardless of the amount of time given to instruction.

A second conclusion based on these results is that either the subtest of Literal Comprehension or the objectives upon which it is based is not appropriate. No group, regardless of ability level or school assignment, achieved a mean score which met the criterion.

Discussion of Inferential Analyses

Hypotheses I and II

Hypotheses I and II were tested using a multiple regression linear model. Significant relationships were hypothesized to exist between scores achieved on the Ginn Level 10 mastery test (and subtests) and overall reading ability, school assignment, and their interaction. The subsamples of above and below average readers were used in this analysis.

The relationship between performance on the Ginn Level 10 mastery test and overall reading ability was significant at the total test level and for all the eight subtests. The null hypothesis was rejected for both Hypotheses I and II. The researcher concluded that the relationship between overall reading ability and performance on the Ginn Level 10 mastery test is significant and overall reading ability is a significant predictor of performance on this program-dependent mastery test.

The variable of school assignment was not a significant predictor of performance on this mastery test at the total

test level. School assignment was a significant predictor, in addition to the variance accounted for by overall ability, to performance on one subtest, Vocabulary I. This result suggests that faculty at certain schools may have developed alternative strategies for teaching vocabulary (word meanings). Performance on this subtest may also be affected by contextual variables which would help to develop vocabulary in students.

The test for an interaction between school assignment and overall ability was significant for one subtest, Decoding I. This interaction was significant in addition to the variance accounted for by school assignment and overall ability. This interaction indicates that performance on this subtest was differentially affected by school membership for certain levels of overall ability.

To ascertain the nature of the interaction, the regression coefficients for each school were calculated and the regression lines, by school, were graphed. This graph is presented in Figure 1. Criterion score to pass this subtest was four. Two schools (school 2 and school 9) observed scores where students with even lowest levels of overall ability were above this criterion. In these schools, the difference in observed scores between lower levels and higher levels of ability was minimal. Overall ability was probably not a significant predictor for performance on this subtest within these schools. In addition, below average

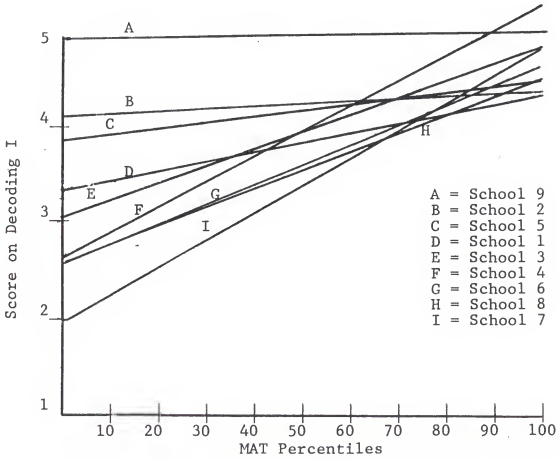

Figure 1. Interaction, by School, Between Scores on Decoding I Subtest and MAT

and above average readers in schools 9, 2, 5, and 1 not only had similar levels of achievement but their achievement was uniformly high. These results may indicate that for certain skills at least, the usual observed relationship between overall ability and performance can be minimized and that some instructional practice in these schools has been effective in reducing the gap in performance between above average and below average readers. In contrast, in schools 3, 4, 6, 7, and 8, lower levels of performance were observed, particularly for below average students and the visibly steeper slopes depict a positive relationship between overall ability and performance on the mastery test. Further research is needed to ascertain what factors within schools are effective in minimizing this relationship.

Hypothesis III

Hypothesis III tested for a significant difference in the proportion of above average and below average readers who attained mastery level status on the Ginn Level 10 mastery test. The results of this analysis indicate that for this population a significantly higher proportion of above average readers attained mastery of Ginn Level 10. Ninety-two percent of the above average readers achieved mastery while only 38 percent of the below average readers attained mastery.

The result of Hypothesis III is crucial to the central question of this study. Attaining a mastery level criterion is the objective of a mastery learning model. The results of the analysis of hypothesis III prove that the below average readers in this population do not attain mastery as frequently regardless of the differential time they are given instruction.

Discussion of Item Analyses

Performance on each of the ninety-one items on the Ginn Level 10 mastery test was compared for above average and below average readers. Comparisons were made for three item parameters: item p-values, latent trait difficulty indices, and goodness-of-fit to the one parameter logistic model. The results of these computations were presented in Table 14.

Item p-Values

The results of the item p-value analyses indicate that as a group the below average readers experienced greater difficulty on a majority of the items on the Ginn Level 10 mastery test. In addition, this group had a greater range of item difficulties. Only 33 percent of the items were passed by 80 percent of the below average readers compared to 80 percent of the items passed at this level by above average readers. Since mastery criteria are set at 80 percent, this observation is important. Haladyna and Roid

(1981a) argued that criterion-referenced tests could be more accurate if item difficulties were matched to ability levels of students being tested. For this population of students, the item difficulties seemed to match the ability levels of the above average readers. The difficulties did not match the ability levels of the below average readers.

Latent Trait Difficulty Indices

The item difficulties computed for the above average and below average readers by the latent trait model would be expected to differ. However, the difference in difficulty between these groups should remain relatively constant if the item is testing the same trait for each group of students. The item difficulties were correlated twice: once for the total sample of items (n=91) and again for this sample minus items which misfit the model (n=77).

The correlations between all items were significant at alpha level .01. The correlation equaled .735. When items which misfit the latent trait model were dropped the correlation increased slightly to .744. This correlation suggests that relative item difficulties remained fairly consistent between the above average and below average readers. There is little evidence to suggest different traits are being measured for these two groups. However, this evidence does not prove the trait being measured is only that which has been taught through Ginn Level 10 instruction.

Fit to the One Parameter Logistic Model

Of the ninety-one items used on the Ginn Level 10 mastery test, fourteen items significantly misfit the one parameter logistic model according to Wright's (1977) criterion for significance. An item misfitting the model is not measuring the trait being tested in the same way for all students in the group. Of the fourteen items misfitting the model, ten items misfit the model generated for the below average readers and four items misfit the model generated for the above average readers. This result suggests that particular items, and possible particular subtests are measuring separate traits.

The fourteen items which did not fit the one parameter model were reviewed. A structural review focused on item format and test directions. A content review focused on the instructional objectives to which each item was keyed. Table 17 presents all item statistics which were calculated for these fourteen items along with a description of the skill each tested. Of these fourteen items, eleven tested comprehension skills. Three of these tested literal comprehension and eight tested inferential comprehension skills. These results indicate that about one third of the total questions on comprehension were in some way inadequately testing the skill desired for at least part of the student group being tested.

Table 17

Items from the Ginn Level 10 Mastery Test Which Did Not Fit the One Parameter Logistic Model for Above Average or Below Average Readers

Subtest/Item	Subgroup*	Skill Tested	p-Values Below Average Readers	p-Values Above Average Readers
Comprehension I Item 1	Below Average Readers	Details	.70	.92
Comprehension I Item 2	Below Average Readers	Main Idea	.70	.77
Comprehension I Item 8	Below Average Readers	Main Idea	.58	.60
Comprehension II Item 3	Above Average Readers	Main Idea	.75	.75
Comprehension II Item 7	Below Average Readers	Predicting Outcomes	.70	.93
Comprehension II Item 8	Below Average Readers	Predicting Outcomes	.80	.97
Comprehension II Item 10	Below Average Readers	Inferring Character Trait	.72	.94
Comprehension II Item 11	Above Average Readers	Inferring Fact/Opinion	.68	.84

Continued

Table 17

Continued

Subtest/Item	Subgroup*	Skill Tested	p-Values Below Average Readers	p-Values Above Average Readers
Comprehension II Item 12	Below Average Readers	Inferring Cause/Effect	.89	.83
Comprehension II Item 16	Above Average Readers	Details	.70	.80
Comprehension II Item 18	Below Average Readers	Predicting Outcomes	.44	.46
Decoding I Item 4	Below Average Readers	4-Syllable Words	.46	.69
Decoding II Item 4	Above Average Readers	Digraphs ci as in ancient	.63	.78
Decoding III Item 3	Below Average Readers	Vowels y as in gym	.63	.84

*This column refers to the subgroup for which the item did not fit the one parameter logistic model.

Structural Review of Items

Of the eleven misfitting comprehension items, five had formats which differed from the majority of items on the test. For two items, students were asked to circle more than one answer (two answers for one, three for the other) although the item was still dichotomously scored. For two other items, students were asked to read an underlined statement, then infer an answer to a question based on that statement. The fifth item with a variation in format first asked students to separate certain paragraphs from the total story, then pick an answer based only on their interpretation of those paragraphs. These variations in format were mixed throughout the entire set of questions.

Test directions for all misfitting comprehension items were student read. Once the test administrator gave overall instructions to begin the test, each student was expected to read each question individually, interpret the question, and find the correct answer(s).

Content Review of Items

Six of the misfitting comprehension items did not have irregular formats. Three of these tested the skill of selecting the main idea of a passage. On this mastery test, eight items were keyed to this skill. Therefore, almost one third of the items testing this skill misfit the one parameter logistic model and may be inaccurately testing this

skill. Three other items tested the inferential skill of predicting outcomes. These items represented three-fourths of the items keyed to this objective. These results suggest either that these skills may not be adequately developed and taught or that they are not being accurately tested. For below average readers, these inferential skills may require a knowledge base not yet developed to the abstract level required on this test.

The three decoding items misfitting the model may be due to student deficits in vocabulary or contextual reading more than to inability to decode. Two tasks were required to correctly answer each question. First the student had to correctly decode each foil. Then the student had to choose an answer which fit the sentence context. The p-values for these items were lower than the other items in these subtests. Students may have been decoding properly, however they may not have understood the word meaning.

Conclusions

This study was designed to determine to what extent pupil performance on a program-dependent mastery test is determined by overall reading ability, school assignment and their interaction. The results of the analyses show that overall reading ability was significantly related to performance on the program-dependent mastery test and eight subtests at the .05 level of significance. School assignment was related significantly (.01 alpha level) to performance

on the mastery test for one subtest; one interaction was significant at alpha level .025. The proportion of above average readers attaining mastery criterion was significantly higher than the proportion of below average readers at alpha level .01.

Based on the results of this study, the researcher concludes that for this population and this mastery test, the average performance of below average readers was significantly lower than the performance of readers with above average ability. Below average readers had greater difficulty at the total test level, subtest level and item level than the above average readers. In addition below average readers had greater difficulty than above average readers in reaching the mastery criteria set for passing the total test and most subtests.

The results of this study could be caused by numerous factors. First, a possibility exists that the mastery model of instruction is not being practiced by all teachers in all schools. For this sample however, the researcher is certain the six components are present in each school and being monitored by specialized personnel. Second, all components within this competency-based reading program may not be adequate. If so, feedback to teachers from each of these components may be inaccurate causing planning for students to be improper. Third, the test may not be measuring all skills accurately. The item analyses indicate some item deficiencies may exist on this mastery test. And fourth,

the possibility exists that differential time in treatment, although important, may not be sufficient for helping below average readers attain mastery of any reading level. Differential teaching methods, materials, and possibly a differential ordering of objectives may be necessary for helping students with lower aptitudes in reading. The interaction noted in hypothesis II suggests that some interventions for this group did work in a few schools. A need still exists to ascertain what types of interventions worked.

Recommendations

Implications for Pedagogical Practice

1. Competency-based learning programs can be used with confidence only when evidence is available that all components have been evaluated.
2. When reviewing a competency-based program, district level departments of evaluation should be included on a review team to help ascertain if the program and tests, have been properly validated. If they have not, the program should not be used.
3. Schools should not depend on testing alone to make decisions about student achievement.
4. Teachers cannot depend upon a basic program even if competency-based, to meet the instructional needs of students below average in ability.

5. When using a competency-based system of instruction, equal emphasis should be placed on each of the six components.

Recommendations for Further Research

1. Research is needed to determine the extent to which changes in item format affect test results for elementary students.
2. Research is needed on test administration practices for criterion-referenced tests at elementary level.
3. Research is needed to establish guidelines for evaluating all the components of a competency-based program.
4. Research is needed to ascertain what variables may be causing the interaction between overall reading ability and school assignment. These variables should be classified into contextual, teacher-related, objectives-related or materials-related and later experimentally researched.

REFERENCES

Anastasi, A. Psychological testing (4th ed.). New York: Macmillan, 1976

Berk, R.A. Criterion referenced measurement: The state of the art. Baltimore: Johns Hopkins University Press, 1978.

Berk, R.A. A consumers' guide to criterion referenced test reliability. Journal of Educational Measurement, 1980, 17, 323-349.

Bloom, B.S. Mastery learning and its implications for curriculum development. In Elliot W. Eisner (ed.), Confronting curriculum reform. Boston: Little, Brown, 1971.

Bloom, B.S. Human characteristics and school learning. New York: McGraw-Hill, 1976.

Block, J.H. Mastery learning: Theory and practice. New York: Holt, Rinehart and Winston, 1971.

Block, J.H., and Anderson, L.W. Mastery learning in classroom instruction. New York: Macmillian, 1975.

Bobbitt, J.F. The curriculum. New York: Houghton Miflin, 1918.

Brennan, R.L., and Kane, M.T. An index of dependability for mastery tests. Journal of Educational Measurement, 1977, 14, 277-289.

Brennan, R.L., and Stolurow, L.M. An empirical decision process for formative evaluation. Research Memorandum No. 4. Cambridge, MS: Harvard University CAI Laboratory, 1971.

Brittain, M.M., Domain referenced testing of reading achievement. Paper presented at the annual meeting of the Eastern Educational Research Association, Philadelphia, March 1981. (Eric Document Reproduction Service No. ED 210 621)

Carroll, J.B. A model of school learning. Teacher's College Record, 1963, 65, 723-733.

Carroll, J.B. Importance of the time factor in learning. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, 1973.

Charters, W.W. Curriculum construction. New York: Macmillan, 1923.

Clymer, T., Blanton, W., Johnson, D., and Lapp, D. A lizard to start with (Level 10) mastery test. Reading 720. Ginn and Company (Xerox Corporation): Lexington, Mass., 1976.

Cox, R.C., and Vargas, J.S. A comparison of item selection techniques for norm-referenced and criterion-referenced tests. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, April 1966.

Cramer, E.M., and Appelbaum, M.F. The validity of polynomial regression in the random regression model. Review of Educational Research, 1978, 48, 511-515.

Cronbach, L.J. Comments on mastery learning and its implications for curriculum development. In Elliot W. Eisner (ed.), Confronting curriculum reform. Boston: Little, Brown, 1971.

Cronbach, L.J., Gleser, G.C., Nanda, H., and Rajaratnam N. The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York: Wiley, 1972.

Drahozal, F,C, and Hanna, J.C. Reading comprehension subscores: Pretty bottles for ordinary wine. Journal of Reading, 1978, 21, 416-420.

Ebel, R.L. Content standard test scores. Educational and Psychological Measurement, 1962, 22, 15-25.

Flanagan, J.C. Units, scores, and norms. In E.T. Lindquist (ed.) Educational measurement. Washington, D.C.: American Council on Education, 1951.

Garcia-Quintana, R.A. Person fit to the RASCH model using norm-referenced and criterion-referenced statewide data. Paper presented to the annual meeting of the Southeastern Psychological Association, Atlanta, March 1981. (Eric Document Reproduction Service No. ED 210 307)

Glaser, R. Instructional technology and the measurement of learning outcomes: Some questions. American Psychologist, 1963, 18, 519-521.

Glass, G.V. Minimum competence and incompetence in Florida. Phi Delta Kappan, 1978, 49, 602-605.

Haladyna, T.M. Effects of different samples on item and test characteristics of criterion-referenced tests. Journal of Educational Measurement, 1974, 11, 93-100.

Haladyna, T., and Roid, G.A. A comparison of two item selection procedures for building criterion referenced tests. Washington, D.C: National Institute of Education, 1981(a).

Haladyna, T., and Roid, G. The role of instructional sensitivity in the empirical review of criterion referenced test items. Journal of Educational Measurement, 1981, 18, 39-53(b).

Hambleton, R. Test score validity and standard setting methods. In Berk, R.A. (ed.) Criterion referenced measurement: The state of the art. Baltimore: Johns Hopkins, 1978.

Hambleton, R.K., and Cook, L.L. Latent trait models and their use in the analysis of educational test data. Journal of Educational Measurement, 1977, 14, 75-96.

Hambleton, R.K., and Eignor, D.R. Guidelines for evaluating criterion referenced tests and test manuals. Journal of Educational Measurement, 1978, 15, 321-327.

Hambleton, R.K., and Novick, M.R. Toward an integration of theory and method for criterion-referenced tests. Journal of Educational Measurement, 1973, 10, 159-170.

Hambleton, R.K., Swaminathan, H., Algina, J., and Coulson, D.B. Criterion-referenced testing and measurement: A review of technical issues and developments. Review of Educational Research, 1978, 48, 1-47.

Harnisch, D.L. Analysis of item response patterns: Consistency indices and their application to criterion referenced testing. Paper presented to the Annual meeting of the American Educational Research Association, Los Angeles, April 1981. (Eric Document Reproduction Service No. 209 335)

Harris, M.L., and Stewart, D.M. Application of classical strategies to criterion referenced test construction: An example. Paper presented to the Annual meeting of the American Educational Research Association, New York, 1971.

Helmstadler, G.C. A comparison of Bayesian and traditional indexes of test item effectiveness. Paper presented at the meeting of the National Council on Measurement in Education, Chicago, April 1974.

Horton, L. Mastery learning. Bloomington, Indiana: Phi Delta Kappa Educational Foundation, 1981.

Huynh, H. On the reliability of decisions in domain referenced testing. Journal of Educational Measurement, 1976, 13, 253-264.

Linn. R.L. Issues of reliability in measurement for competency based programs. In Bunda, M., and Sanders, J. (eds.) Issues in competency based measurement. Washington, D.C.: National Council of Measurement in Education, 1979.

Livingston, R. Criterion referenced applications of classical test theory. Journal of Educational Measurement, 1972, 9, 13-26.

Lord, F.M., and Novick, M.R. Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968.

McClung, M.S. Competency testing: Potential for discrimination. Clearing House Review, 1977, 51, 439-448.

Mean, R., Wright, B., and Bell, S. BICAL - Version 3. Chicago: University of Chicago, 1979.

Mehrens, W.A., and Lehmann, I.J. Measurement and evaluation in education and psychology. New York: Holt, 1969.

Millman, J. Criterion referenced measurement. In Popham, W.J. (ed.) Evaluation in education: Current applications. Berkeley, California: McCutchan, 1974.

Millman, J. Determinants of Item Difficult. Center for the Study of Evaluation Report No. 114. Los Angeles: California University, 1978. (Eric Document Reproduction Service No. 163 071)

Popham, W.J., and Husek, T.R. Implications of criterion referenced measurement. Journal of Educational Measurement, 1969, 6, 1-9.

Popham, W.J. Indices of adequacy for criterion referenced tests. In W.J. Popham (ed.) Criterion referenced measurement: An introduction. Englewood Cliffs, New Jersey: Prentice-Hall, 1971.

Popham, W.J. Criterion referenced measurement. Englewood Cliffs, New Jersey: Prentice-Hall, 1978.

Prescott, G., Balow, I., Hogan, T., and Farr, R. Metropolitan achievement test. New York: The Psychological Corporation, 1978.

Shepard, L., Camilli, G., and Averill, M. Comparison of procedures for detecting test item bias with both internal and external ability criteria. Journal of Educational Statistics, 1981, 6, 317-375.

Shoemaker, S.H., and Johnson, R.T. Assessing the construct validity of a criterion referenced test: A nemological network approach. A paper presented at the Annual Meeting of the American Educational Research Association, April, 1981. (Eric Document Reproduction Service No. 204 359).

Shuy, R.W. The reading teacher's knowledge matters more than texts or tests. The Education Digest, 1982, 47, 54-47.

Smith, D. The effects of various item selection methods on the classification accuracy and classification consistency of criterion referenced instruments. A paper presented at the Annual Meeting of the American Educational Research Association, Toronoto, Ontario, Canada, March, 1978. (Eric Document Reproduction Service No. 159 222).

Skager, R.W. The great criterion referenced test myth. Paper presented at the Annual Meeting of the American Educational Research Association, Washington, D.C., April, 1975. (Eric Document Reproduction Service No. 160 650).

Swaminathan, H., Hambleton, R., Algina, J. Reliability of criterion referenced tests: A decision theoretic formulation. Journal of Educational Measurement, 1974, 11, 263-267.

Subkoviak, M.J. Estimating reliability from a single administration of a mastery test. Journal of Educational Measurement, 1976, 13, 265-276.

Tatsuoka, M., and Tatsuoka K. Detection of aberrant response patterns and their effect on dimensionality. Computer based Education Resource Laboratory Research Report No. 80-4. Urbana: University of Illinois, 1980.

Thorndike, E.L. Educational psychology, Vol. I. New York: Teacher's College, Columbia University, 1913.

Torshen, K.P. The mastery approach to competency based education. New York: Academic Press, 1977.

Tyler, R.W. Basic principles of curriculum and instruction. Chicago: University of Chicago Press, 1950.

Walmsley, S.A. The criterion referenced measurement of an early reading behavior. Reading Research Quarterly, 1979, 14, 574-603.

Wright, B.D. Solving measurement problems with the RASCH model. Journal of Educational Measurement, 1977, 14, 97-116.

BIOGRAPHICAL SKETCH

Darla McCrea was born in Kalamazoo, Michigan, in 1951. She attended public schools and graduated from Parchment High School in 1969. During her senior year she worked as a teaching assistant in a developmental kindergarten class.

In 1972, Darla received a B.A. degree from Western Michigan University with specializations in elementary education, sociology and Asian studies. While at W.M.U., Darla was a member of the Honors College and was awarded a Waldo-Sangren Award for a project working with kindergarten children.

Between 1972 and 1976, Darla lived in Charlottesville, Virginia. She received a master's degree from the University of Virginia in 1975 in Social Foundations of Education.

In 1976, Darla moved to Gainesville, Florida. She entered University of Florida's College of Education, Department of Instructional Leadership and Support, in 1979.

Between the years of 1972 and 1983, Darla taught, in four elementary schools, in grades Kindergarten, First, Second, Third, and Sixth. She worked for four years as a curriculum resource teacher and is currently an elementary principal.

Darla lives in Gainesville, Florida, with her husband, Brian. They have three children: Sara, Sam and Jacob.

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

William D. Hedges, Chairman
Professor of Instructional
Leadership and Support

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Gordon D. Lawrence
Professor of Instructional
Leadership and Support

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Linda M. Crocker
Associate Professor of
Foundations of Education

This dissertation was submitted to the Graduate Faculty of the Division of Curriculum and Instruction in the College of Education and to the Graduate School, and was accepted as partial fulfillment of the requirements for the degree of Doctor of Philosophy.

December 1983

Dean for Graduate Studies and Research